Method for Inferring the Optimal Number of Clusters with Subsequent Automatic Data Labeling based on Standard Deviation

Aline Montenegro Leal Silva[1], Francisco Alysson da Silva Sousa[2], Alysson Ramires de Freitas Santos[3],
Vinicius Ponte Machado[4], André Macedo Santana[5]
Federal University of Piaui[1,2,4,5], Brazil
Unified Teaching Center of Piaui[3], Brazil

***Abstract*—Machine learning is a suitable pattern recognition technique for detecting correlations between data. In the case of unsupervised learning, the groups formed from these correlations can receive a label, which consists of describing them in terms of their most relevant attributes and their respective ranges of values so that they are understood automatically. In this research work, this process is called labeling. However, a challenge for researchers is establishing the optimal number of clusters that best represent the underlying structure of the data subjected to clustering. This optimal number may vary depending on the data set and the grouping method used and influences the data clustering process and, consequently, the interpretability of the generated groups. Therefore, this research aims to provide an inference approach to the number of clusters to be used in the grouping based on the range of attribute values, followed by automatic data labeling based on the standard deviation to maximize the understanding of the groups obtained. This methodology was applied to four databases. The results show that it contributes to the interpretation of the groups since it generates more accurate labels without any overlap between ranges of values, considering the same attribute in different groups.**

***Keywords*—Inference approach; range of attribute values; labeling; standard deviation; interpretation of the groups**

I. Introduction

The rapid popularization of computers in many sectors of society has resulted in significant data volume growth [1]. The researchers then began to use pattern recognition techniques by detecting correlations between the data, which could bring to light relevant and valuable knowledge potentially contained in these databases [2].

One of these pattern recognition techniques is machine learning (ML), which emerged from the realization of creating computer programs that learn a particular behavior or pattern automatically from examples or observations. The idea behind learning is that after looking at some data, a computer builds a model based on that data and uses that model as a hypothesis about the world and a piece of software that can solve problems [3].

Machine learning can consist of two main paradigms: supervised and unsupervised. For supervised learning, the aim is to create an accurate model for predicting values for new data. As for unsupervised learning, the objective is to find characteristics that can summarize the data. Other paradigms exist, such as reinforcement learning, multitasking, and semi-supervised.

As one of the best-known techniques in the area of unsupervised learning, grouping or clustering of data consists of defining a set of groups or clusters in which the elements of the same group are as similar as possible to each other, and the elements of different groups are as distinct as possible [4]. Because it is subjective, it does not provide clear information that allows inferring the characteristics of each cluster formed [5] due to the algorithms' limitations.

Establishing the optimal number of groups in a clustering algorithm is one of the most challenging and fundamental tasks for researchers since different amounts of clusters cause different results, influencing the performance of the clustering process [6], [7], [8]. For example, Cobweb [9] is a hierarchical algorithm whose order of factors affects the grouping and is very sensitive to data input. K-means [10] is an algorithm based on Euclidean distance dependent on the initial partition generated by the random choice of centroids, requiring the number K of groups to be informed in advance.

Labeling seeks to synthesize its definition, describing the groups' most relevant attributes and respective value ranges to understand the specialist better. Due to some limitations resulting from the grouping, auxiliary techniques can infer characteristics that identify the formed groups. Among these techniques are dispersion metrics, such as standard deviation. Therefore, the resulting clusters are labeled to be understood automatically. In this research work, this process is called automatic data labeling, which aims to identify the characteristics of each group and, later, allow the complete interpretation of the generated clusters.

In this sense, this research aims to provide an inference approach to the number of groups to be used in the clustering process, based on the range of attribute values, with subsequent automatic data labeling from the standard deviation to maximize the understanding of the groups obtained, without overlapping any range of values in the same dataset, considering the same attribute. It consists of calculating the standard deviation according to the value of each attribute. When necessary, the value of this deviation is increased to the lowest value of each attribute in the dataset or decreased by the highest value of each attribute, or both steps have been performed. This methodology was applied to four databases of different sizes.

The results show that it contributes to the interpretation of groups, as it generates more accurate labels since the algorithm developed to choose the optimal number of groups performs better when compared to other isolated methods, such as Elbow, Silhouette Coefficient, and Calinski-Harabasz, for example. Furthermore, the labels obtained do not overlap between ranges of values considering the same attribute in different groups. It contributes to better interpretability of the generated clusters.

In addition to the Introduction, the rest of this article is organized as follows. Section II presents the theoretical framework used in this model, Section III addresses influence studies for this research, Section IV displays the methodology used, Section V presents the results obtained and, finally, Section VI describes the conclusion of the research work.

II. Theoretical Reference

In this session, the selection processes of relevant attributes and measures of dispersion of data used in this research will be presented, followed by clustering, in addition to the labeling problem.

A. Dispersion Measures

According to [10], a measure of dispersion for a quantitative variable indicates the degree of spread of sample values around the centrality measure, indicating how much the elements differ from the mean of the data set. Greater dispersions exhibit less representativeness of the central values. One of the advantages of using these metrics is that they observe the data set as a whole and assess the degree of homogeneity or dispersion of this set, favoring a reduced computational cost. Next, a relevant metric used in data labeling will be presented to understand the *clusters* formed.

1) Standard Deviation: Mathematically, the standard deviation (SD) is a measure of dispersion used to quantify the variation or dispersion of a set of data values [11]. Equation 2 shows how the standard deviation calculation is performed.

$$DP = \sqrt{\frac{\sum\limits_{i=1}^{n}(x_{\mathrm{i}} - MA)^2}{n}} \tag{1}$$

For clarification:

- SD: standard deviation;
- $\mathrm{x_i}$: value at position i in the data set;
- MA: arithmetic mean of the data;
- n: the amount of data.

A low standard deviation means that the data points tend to be close to the mean of the set, while a high standard deviation indicates that the data points are spread over a wide range of values.

B. Clustering

The basic idea of clustering is those elements that make up the same *cluster* must show high similarity (i.e. be very similar elements and follow a similar pattern). Still, it must be very dissimilar from objects in other groups. In other words, all clustering is done to maximize homogeneity within each *cluster* and maximize heterogeneity between groups.

K-means is one of the most popular clustering algorithms. The result of the K-means [12] method is generally influenced by the K-partition chosen in the initial step. If K is too small, there will be distinct elements in the same cluster. On the other hand, if K is very high, similar elements will be in different clusters. For this reason, it is recommended to validate the result of the cluster analysis based on inference criteria of the optimal number of groups in a data set, bearing in mind that different amounts of *clusters* generate different results, influencing the performance of clustering and consequently in understanding the *clusters* formed.

C. Labeling Problem

The task of interpreting clusters is commonly assigned to a specialist in the field under study who examines each group with respect to its objects to label them, describing the nature of the group. This process tends to be too laborious concerning time and resources, considering the amount of data and subjectivity of the task.

In view of this, [14] proposed a method for automatic extraction of characteristics from the groups, providing specialists a label with a selection of the most relevant characteristics of the elements of each group. These features are composed of attribute values range, so the labeling problem is defined as:

Given a set of clusters C =$\{\mathrm{c_1}, ..., \mathrm{c_k} \mid \mathrm{k}{\geq}1\}$, so that each cluster contains a set of elements $\mathrm{c_i} = \{\overrightarrow{e}_1, ..., \overrightarrow{e}_{\mathrm{n(c_i)}} | \mathrm{n}^{(\mathrm{c_i})}{\geq}1\}$ which can be represented by a vector of attributes defined in $\mathrm{R^m}$ and expressed by $\overrightarrow{e}_{\mathrm{j}}{}^{(\mathrm{c_i})}$ = $(\mathrm{a_1}, ..., \mathrm{a_m})$ and even though $c_1 \cap c_{1'} = \emptyset$ with $1 \geq$ i, i '$\geq$ K and i $\neq$i'; it aims to present a set of labels R =$\{\mathrm{r_{c_i}}, ..., \mathrm{r_{c_k}}\}$ in which each specific label is given by a set of pairs of values, attributes and their respective range $r_{(\mathrm{c_i})} = (a_1, [p_1, q_1]), ..., (a_{\mathrm{m(c_i)}},]p_{\mathrm{m(c_i)}}, q_{\mathrm{m(c_i)}}])$ able to better express the associated $\mathrm{c_i}$ cluster.

In order to clarify:

- K is the number of clusters;
- $\mathrm{c_i}$ is any cluster;
- $n^{(c_{\mathrm{i}})}$ is the number of elements in cluster $\mathrm{c_i}$;
- $\overrightarrow{e}_{\mathrm{j}}{}^{(c_{\mathrm{i}})}$ refers to the j-th element belonging to cluster $\mathrm{c_i}$;
- m is the dimension of the problem;
- $\mathrm{r_{(c_i)}}$ is the label for cluster $\mathrm{c_i}$;
- $]\mathrm{p_{m(c_i)}},\mathrm{q_{m(c_i)}}]$ represents the values range of attribute $\mathrm{a_{m(c_i)}}$ where $\mathrm{p_{m(c_i)}}$ is lower limit and $\mathrm{q_{m(c_i)}}$ is upper limit;
- $m^{(c_{\mathrm{i}})}$ is the number of attributes present in a label for cluster $\mathrm{c_i}$.

Finally, the method has as input a set of clusters and must present as output a specific label for each group that best defines it, according to the specifications already presented.

III. Related Work

This section addresses some methods for inferring the number of groups in the data clustering process, as well as automatic cluster labeling models that influenced this research, presenting its methodologies and results obtained.

In [15], a model was proposed that can be applied to the segmentation of products for inventory management based on the analysis of three basic principles, which are: history (recency (R)), frequency (F), and money spent (monetary (M)) from the K-means algorithm. Meanwhile, the determination of the optimal number of clusters was evaluated using eight validation indices, namely, Elbow Method, Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index, Ratkowski Index, Hubert Index, Ball-Hall and Krzanowski-Lai Index to improve objectivity and accuracy in product segmentation compared to using only one method. The result obtained in all these criteria was 3 clusters, the optimal number of groups, with a low variance between the intra-cluster data, resulting in a high similarity between the elements of the same group.

According to [16], a method was presented to classify the egg production of laying hens in Indonesia based on the K-Means clustering algorithm. The survey data was taken from the National Statistics Center of Indonesia and corresponded to the period from 2018 to 2020 from 34 provinces. To validate the number of groups to be used, the researcher evaluated the Davies Bouldin Index (DBI) criterion for each number of existing clusters, which consists of the ratio between the intra-cluster and inter-cluster distances. In this study, 8 clusters were used, and the DBI value was calculated for each. It was observed that the optimal number of groups is four since it has the lowest DBI value.

A model proposed by [14] groups data based on the centroids of the clusters and uses Artificial Neural Networks (ANN) to generate labels for each of them. Initially, a dataset was provided as input to the model. To obtain better performance for continuous values, a discretization process was performed, in which different possible values for each attribute were divided into intervals, which represent the range of values. In the second stage, the clustering process was carried out using an unsupervised algorithm (K-means). Once the groups were generated, a supervised algorithm (ANN) was applied to each of them, using the discretized base to detect which attributes were relevant in the formation of each generated group. This methodology was applied to three databases (Glass, Seeds, and Iris), and the results were obtained with an average more excellent than 88.79% of correctly labeled elements.

The work of [1] used unsupervised and supervised machine learning methods for data clustering and labeling tasks. To group the data, the DAta MIning COde REpository (DAMICORE) algorithm was used, and to label, the Automated Labeling Method (ALM) based on Artificial Neural Networks (ANN) was used. Before data grouping, the data sets were submitted to the discretization step, and the continuous attributes were discretized by the EWD and EFD methods. The results were compared with those presented in the [13] model, and the analysis showed that applying the ALM method generated better results. The groups formed by DAMICORE are more accurate than those obtained by applying the K-Means cluster, with an average accuracy above 90%.

IV. Proposed Method

This research aims to provide an approach for inferring the number of groups to be used in the grouping process, based on a range of attribute values, with subsequent automatic data labeling from standard deviation to maximize the understanding of the groups obtained. Initially, the new method was validated based on a model already proposed in the literature, that of [14], which developed an algorithm for automatic extraction of the characteristics of the groups, contributing to the interpretability of these clusters.

The algorithm K-means in the proposed model used Python's sklearn library, and the ease and robustness of the environment could provide tests that led to a better understanding of the problem addressed.

Table I presents the four databases used, starting from the UCI Repository[1], including Wine, Breast Cancer, Quality White Wine, and Credit Card.

TABLE I. Databases Obtained from the UCI Repository

Databases	Amount of Data	Attributes
Wine	178	13
Breast Cancer	699	10
Wine Quality White	4.898	12
Credit Card	30.000	24

The methodology used by [14] to aid this task of interpreting clusters is illustrated in Fig. 1.

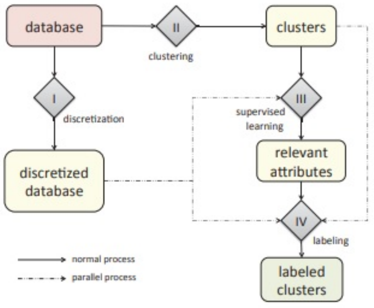

Fig. 1. Labeling template flowchart from [14].

At first, this model receives a database as an input parameter. This base can contain different types of data - discrete or continuous. In some cases, it will be necessary to apply a discretization method (I), which consists of assigning discrete values to attributes that can assume a wide variety of values within a given domain. Thus, the supervised learning algorithm used in step III will be able to identify a possible relationship between attributes with less complexity, showing better results when dealing with the classification problem involving such attributes. According to [17] and [18], there may be an increase in accuracy and speed during the training stage when using a discretization method. In addition, this discretization process allows the inference of a value range, which happens in step IV.

[1] https://archive.ics.uci.edu/ml/index.php

The discretization process starts with selecting which attributes should be discretized and which type to use. Therefore, for this model, the number of ranges of values was defined as FX, and the discretization technique used was by equal frequencies (EFD) to avoid an inadequate distribution of the values of an attribute due to identical elements and consequently cause an imbalance in the distribution of these elements about the ranges of values. In EFD, the number of components with different values between the cut-off points remains constant.

The second step (II) corresponds to using an unsupervised algorithm that receives as input a set of elements (in this case, the database) and presents as output the association of each component to a respective created cluster. The discretized database is not used in this step, but the initially provided database. The algorithm used was K-means, but any other algorithm with unsupervised learning capable of dealing with the clustering problem can be used.

Then step (III), an algorithm with supervised learning is applied to detect the relevant attributes for the definition of each group once the clusters are appropriately formed and the data to be worked on, if necessary, are already discretized. Then, the actual labeling work begins. Each label referring to any group is based on a set of attributes and their respective ranges of values. Therefore, this step has as input a set of clusters and presents as output a set of attributes for each generated group that will be used in its labeling. For this, artificial neural networks of the Multi-Layer Perceptron (MLP) type were used. However, in principle, any other algorithm with supervised learning capable of detecting relationships between variables or any other technique capable of selecting attributes could be chosen. In this case, each neural network presents a hit rate for its learning, performed only with the elements of their respective clusters.

Finally step (IV), a strategy that selects the value (for discrete attributes) or value range (for continuous attributes) for each chosen relevant attribute is applied to generate labels. This strategy seeks to represent the majority of the group so that the selected values for each attribute are those with the highest frequency in the group. The neural network chose the most relevant attributes in step III.

It was noticed, therefore, that the [14] model did not use any inference criteria to optimize the number of groups in the clustering process and did not verify the best amount of range of values for the composition of the labels.

Thus, this research work has the initial intention of providing a method for inferring the optimal number of groups considering different ranges of attribute values, since in the [14] model, as the number of ranges increases of values for the same group K, the hit rate of this model decreases. This fact was found in the four databases used, as shown in Table II.

Based on Table II, this research work developed the method *Optimization of the Number of Clusters Based on a Range of Values* according to the performance analysis (hit rate) of the [14] model to find the optimal number of groups and range of values to be used in the grouping step.

TABLE II. Performance Analysis (Hit Rate (%)) for Variations of K and Value Range (FX) According to the [14] Model

WINE	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**
FX=2	95.51	98.96	97.17	98.84	98.90
FX=3	87.29	90.78	96.06	98.21	98.59
FX=4	78.93	86.23	85.07	91.24	92.71
FX=5	71.41	75.36	82.06	87.22	91.88
FX=6	67.27	70.48	73.16	85.84	87.96

BREAST_CANCER	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**
FX=2	91.34	93.41	92.55	92.04	91.58
FX=3	90.48	86.44	91.05	91.86	90.96
FX=4	86.28	82.17	87.67	91.05	89.95
FX=5	77.39	79.98	75.40	80.01	78.70
FX=6	76.90	74.31	74.50	79.17	78.25

WINE QUALITY WHITE	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**
FX=2	99.90	99.98	100	100	100
FX=3	99.50	99.80	99.94	99.96	99.97
FX=4	98.51	98.66	99.10	99.38	99.61
FX=5	97.24	98.14	97.90	98.38	98.74
FX=6	94.34	94.63	96.06	96.98	97.12

CREDIT CARD	**K=2**	**K=3**	**K=4**	**K=5**	**K=6**
FX=2	92.60	99.94	99.96	99.97	99.98
FX=3	91.50	98.80	98.94	98.90	98.92
FX=4	90.51	97.66	97.10	96.96	97.01
FX=5	87.24	94.14	95.90	95.38	95.12
FX=6	85.34	93.63	94.06	93.98	93.06

1) The starting point of the method was to consider the initial K as the element with the highest frequency among the three inference criteria used (Elbow, Silhouette Coefficient, and Calinski-Harabasz) for the same database, which we call fashion. Table III displays the found values.

   Therefore, the initial K for each of the databases was the following: Wine (3), Breast Cancer (2), Wine Quality White (2), and Credit Card (2).

TABLE III. Criteria for Group Inference

Database	**Elbow**	**Silhouette Coefficient**	**Calinski-Harabasz**
Wine	3	2	3
Breast Cancer	2	2	2
Wine Quality White	2	2	2
Credit Card	2	2	3

2) Next, the range of initial value (FX) equal to the chosen K value was selected, that is, FX = K. For example, the Wine database results will be displayed initially. Therefore, K=3 and FX=3;
3) The model's hits rates for K-1 to K+1 were shown, according to Table IV.

TABLE IV. Hit Rate (%) for Wine Base with K between K-1 and K+1

WINE	K=2	K=3	K=4
FX=2			
FX=3	87.29	90.78	96.06
FX=4			

4) The highest hit rate among those displayed in the previous step was verified. In this case, the highest rate is 96.06% for K=4 and FX=3.
5) Next, the model's hit rate for FX-1 and FX+1 were presented, as shown in Table V.
6) Finally, there was the highest hit rate among the last ones calculated. In this case, for K=4 and FX=2, it was 97.17% hit, and for K=4 and FX=4, it corresponded to 85.07%. Therefore, for this database, the

TABLE V. Hit Rate (%) for Wine Base with FX between FX-1 and FX+1

WINE	K=2	K=3	K=4
FX=2			97.17
FX=3	87.29	90.78	96.06
FX=4			85.07

optimal number of groups and range of values would be K=4 and FX=2 since it was the one that presented the best hit rate. However, in this [13] model, it was found that for this value of K and FX considered optimal, there was partial or complete overlapping of labels in at least two different groups of the same base, considering the same attribute. Therefore, this hit rate is disregarded when such situations occur, and the immediately lower one calculated so far is considered. Therefore, the optimal number of groups and hit rate for the Wine database becomes K=4 and FX=3. Table VI presents what this overlapping is.

TABLE VI. Result of the Group Labeling of the [14] Model for the Wine Database with K=4 and FX=2

Grupo	Elementos	Rótulo		Análise
		Atributos	Faixa	Êxito (%)
0	66	Proline	276.6∼979.0	100
1	23	Alcohol Malic.Acid OD Proline	12.93∼14.83 0.73∼3.27 2.63∼4.0 979.0∼1680.0	100
2	57	Proline	276.6∼979.0	100
3	32	Malic.acid	0.73∼3.27	88.70

According to Table VI, for the [14] model, considering the Wine base with K=4 and FX=2, the following overlaps were found.

- r_{c_0} = (Proline, [276.6∼979.0]) e r_{c_2} = (Proline, [276.6∼979.0]);
- r_{c_1} = (Malic.Acid, [0.73∼3.27]) e r_{c_3} = (Malic.Acid, [0.73∼3.27]).

It means the proline attribute, in *clusters* 0 and 2, and the malic.acid attribute, in *clusters* 1 and 3, have the same range of values. Therefore, these values overlap and no longer represent a single group, making it difficult to interpret the label.

This method was performed for four databases of different sizes and quantities of attributes. It was found that the optimal number of K groups varies between the mode (referring to values from the Elbow, Silhouette, and Calinski-Harabasz methods) and the mode+1. The range of values varies from K-2 to K-1, according to Table VII. The Algorithm 1 summarizes this proposal in pseudocode.

TABLE VII. Optimal Number of Groups and Range of Values

	Optimal Number	
Database	Clusters	Range of Values
Wine	4	3
Breast Cancer	3	2
Wine Quality White	3	2
Credit Card	3	2

Algoritmo 1: Method for Optimizing the Number of Clusters based on the Range of Attribute Values

```
Select initial K by the mode of group inference criteria (Elbow, Silhouette, and
  Calinski-Harabasz);
Make FX = K;
Display the average hit rate of [14]'s method for K between K-1 and K+1;
Display the average method hit rate of [14] for FX between FX-1 and FX+1;
Consider the pair (K, FX) with the highest hit rate of the method among those
  displayed;
while not finding optimal K and FX do
    if complete overlapping of labels in at least two groups of the same base
      then
        Discard the pair (K, FX);
        Search for the next pair (K, FX) whose hit rate is the second highest
    end
    else
        Set the value of K and FX, whose method hit rate is the highest.
    end
end
until Até encontrar K e FX ótimos.
```

Therefore, to develop a more optimized model about the interpretability of the group hit rate and the specificity of the generated labels, considering that no [14] model was found to overlap between ranges of values, assuming the same attribute in different groups, this research work presented a method based on dispersion metrics to solve this limitation of the overlap between ranges of values. The methodology used in this research work went through the following steps, as shown in Fig. 2:

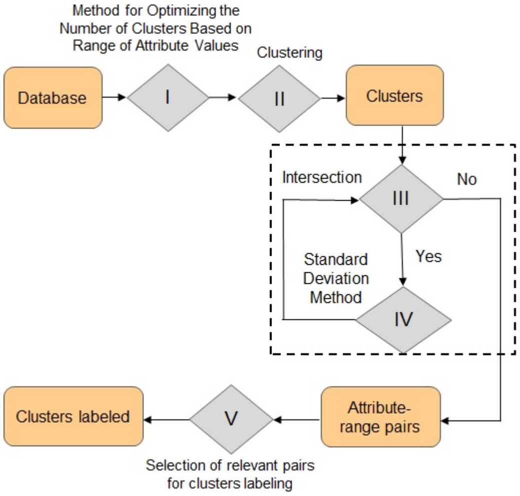
Method for Optimizing the Number of Clusters Based on Range of Attribute Values
Clustering
Database
I
II
Clusters
Intersection
III
No
Yes
Standard Deviation Method
IV
Clusters labeled
V
Attribute-range pairs
Selection of relevant pairs for clusters labeling

Fig. 2. Flowchart of the proposed labeling model.

A. Step 1 - Using the Value Range-Based Cluster Quantity Optimization Method

This step consists of using the proposed method based on the mode of the criteria found in the literature (Elbow, Silhouette Coefficient, and Calinski-Harabasz criterion) for the same data set to find the optimal number of clusters based on a range of values to be used in the data grouping and applied to four original databases of different sizes and quantities of attributes (Wine, Breast Cancer, Wine Quality White, and Credit Card), whose result can be seen in Table VII.

B. Step 2 - Clustering Data

After using the proposed method, the next step corresponds to the grouping of the data, which consists of submitting the original database composed of unlabeled examples to an algorithmic solution of unsupervised machine learning for the formation of groups. The basic idea is that elements that comprise the same group must present high similarity but are very dissimilar from objects in other clusters. The K-means algorithm was used for clustering, but any different clustering algorithm can be used.

C. Step 3 - Standard Deviation Method

Mathematically, standard deviation (SD) is a dispersion measure that is used to quantify the amount of variation or dispersion of a set of data values [11]. A low standard deviation value signifies that data points tend to be close to the mean of the set, while a high standard deviation indicates that data points are spread out over a wide range. Equations 2 show how standard deviation calculation is performed.

$$SD = \sqrt{\frac{\sum_{i=1}^{n}(x_{\mathrm{i}} - M_{\mathrm{A}})^2}{n}} \tag{2}$$

For clarification:

- x_i: value at the i position in the dataset.
- M_A: arithmetic mean of the data.
- n: the amount of data.

Based on the use of standard deviation, a formal definition of the method for automatic cluster labeling is presented below:

Given a set of clusters C_1, C_2, C_3...C_k and A_1, A_2, A_3...A_n the attributes of this subset, the attribute's values distribution A_1 in group C_1 was represented by $A_1C_1(v_i, v_j)$, where i indicates the minimum value and j the maximum value. The technique updates values by applying v_i+SD, v_j-SD, when necessary, considering as a condition exists or not of an intersection between values that A_1 represents, observing the other groups.

Thus, after grouping performed by K-means, attribute values, observing their representation in each group, are organized into temporary structures in which indexes corresponding to the lowest and highest value are identified (Table VIII). Values contained in this table are random for purposes of the understanding method.

TABLE VIII. Indexing of Attribute Values

Attribute Values	8	5	10	6	7
Indexes	0	1	2	3	4
		min.	max.		
		Range: 5 ∼10			

At that moment, it was necessary to check possible overlapping values range in the same dataset, considering the same attribute.

D. Step 4 - Intersection Check

For each attribute, the existence or not of an intersection between values range is verified. If any overlap is found, the standard deviation of the segment under analysis is then calculated based on the arithmetic mean of distances between smallest and largest values, as seen in Table VIII. The updating of these values that identify the ends is applied, observing the need to increase the lowest value or decrease the highest value or even both procedures that, when performed, count as interactions. This process is carried out until there is no longer any intersection between the labels, considering the same attribute in all clusters in the dataset. Fig. 3 contains a representation of values the update applies to until they no longer overlap.

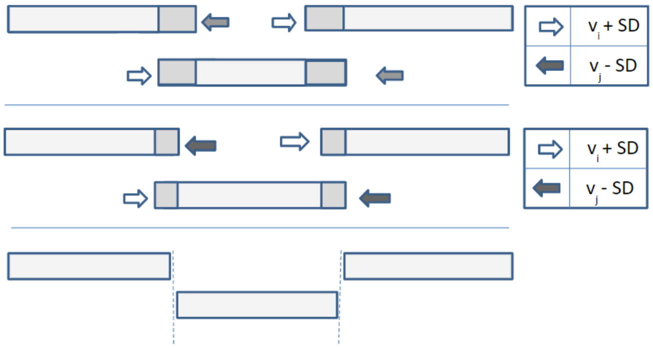

Fig. 3. Sequence of proposed updates.

Considering, for example, the $Petal\_length$ attribute of the Iris dataset, Table IX presents initial values range that represents this characteristic in different groups. In column 1 of this table, groups are identified, column 2 shows the range obtained with grouping, and in columns 3 and 4, operations must be performed when the intersection is found. For this purpose, the standard deviation value found in the respective distribution is applied as an updating factor.

TABLE IX. First SD Iteration for the $Petal\_length$ Attribute of the Iris Dataset

Clusters	Range	Decrement	Increment	SD
C0	3.0 ∼ 5.1	true	false	0.5
C1	1.0 ∼ 1.9	false	false	0.17
C2	4.9 ∼ 6.9	false	true	0.48

This same scenario is also illustrated in Fig. 4. Note that the analyzed attribute overlapped when considering its ranges, which justifies the decrement and increment in $C0$ and $C2$, respectively.

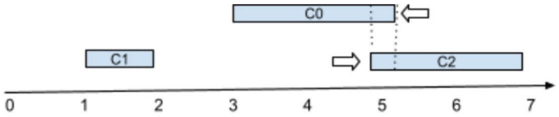

Fig. 4. Scenario described in table IV.

After performing this first iteration, Table X presents the new values resulting from the previous necessary update, also shown in Fig. 5.

TABLE X. SECOND SD ITERATION FOR THE $Petal\_length$ ATTRIBUTE OF THE IRIS DATASET

Clusters	Range	Decrement	Increment	SD
C0	3.0 ∼ 5.0	true	false	0.46
C1	1.0 ∼ 1.9	false	false	0.17
C2	5.0 ∼ 6.9	false	false	0.47

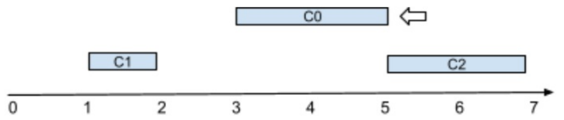

Fig. 5. Scenario described in table V.

Table XI presents data from this update process and verification of possible coincidences in interval segments. As described, it appears that the representation of attribute is distinct in observation of groups formed, thus dispensing with additional iterations. This process can be seen in Fig. 6.

TABLE XI. FINAL RESULT OF THE SD ITERATION FOR THE $Petal\_length$ ATTRIBUTE OF THE IRIS DATASET

Clusters	Range	Decrement	Increment	SD
C0	3.0 ∼ 4.9	false	false	2
C1	1.0 ∼ 1.9	false	false	0
C2	5.0 ∼ 6.9	false	false	1

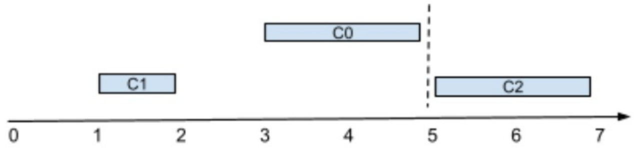

Fig. 6. Scenario described in Table VI.

Algorithm 2 summarizes this Cluster Labeling Proposal in pseudocode, where n is the number of attributes in the dataset, k is the number of groups, and C corresponds to a cluster.

Algoritmo 2: Pseudocode of Cluster Labeling Model Proposed

```
Input: K clusters
for attrA $\leftarrow$ 1 until n do
    while there is an intersection of $att_A$ between two groups do
        $C_i \leftarrow 1$
        for $C_j \leftarrow 2$ until k do
            if max. of attrA in $C_i \geq$ min. of attrA in $C_j$ and max. of attrA in
              $C_i \leq$ max. of attrA in $C_j$ then
                decrement max. of $attr_A$ applying standard deviation of
                  $att_A$ distribution
            end
            if if min of attrA in $C_i \geq$ min attrA in $C_j$ and min. of $att_A$ in $C_i$
              $\leq$ max. $att_A$ in $C_j$ then
                increment min. of $att_A$ applying standard deviation of $att_A$
                  distribution
            end
        end
    end
end
```

E. Step 5 - Selection of the Relevant Attribute-Range Pair

This final step of labeling process consists of selecting relevant attribute–range pair for cluster labels composition, ensuring that each label represents specifically and exclusively one cluster. This selection was based on following measure:

1) Correlation Coefficient: Pearson's association [19] reflects direct relationship between two components, i.e. how much variables are associated, and reach out from +1 to - 1. An association of +1 suggests that there is an extraordinary positive direct relationship between elements, while an association of -1 demonstrates that elements have negative relationship. Table XII presents a categorization for Pearson's correlation coefficient values.

TABLE XII. CATEGORIZING FOR PEARSON'S CORRELATION COEFFICIENT VALUES

Correlation Coefficient	Classification
0	Null
0.01 - 0.19	Very weak
0.2 - 0.39	Weak
0.4 - 0.69	Moderate
0.7 - 0.89	Strong
0.9 - 0.99	Very Strong
1	Perfect

Therefore, the pair(s) of attributes that are most positively correlated in each data set are used in the composition of the final labels of the model.

V. RESULTS AND DISCUSSIONS

The results of applying the proposed method are presented with subsequent cluster labeling based on the correlation coefficient for the selection of the most relevant attributes and on the standard deviation metric to improve the specificity of the range of values of each selected attribute to eliminate overlaps between ranges of values, considering the same attribute in different groups.

In addition, the results obtained were compared with other methods proposed in the literature to show that the model guarantees an improvement in the specificity of the labels, reducing the computational effort to generate them.

A. Iris Dataset

Table XIII presents the analysis result for automatic rotation of the Iris database, for K=4 and FX=2 according to the proposed method. A label describes each cluster with a pair of attribute-value ranges, according to Pearson's association [19].

TABLE XIII. ANALYSIS FOR IRIS DATABASE LABELING

Cluster	Elements	Label		Analysis	Iterations
		Attributes	Range	Hits (%)	
0	28	SL	4.3∼6.1	98	2
		SW	2∼3.2	94	6
1	50	PL	0.99∼3.95	100	0
		PW	0.1∼1.3	100	0
2	32	SL	6.1∼7.9	95	5
		PW	1.3∼2.5	100	0
3	40	PL	3.95∼6.9	99	1
		SW	3.2∼4.4	94	6

Iterations correspond to steps to remove, when necessary, an intersection between clusters in the same dataset, considering the same attribute. It means that the number of iterations of the proposed method for each attribute can differ in each cluster, as it will depend on the dispersion of the group's elements about the mean of the data set. For cluster 0 and the SL attribute, for example, two iterations were necessary to obtain the result, while six iterations for the SW attribute of the same cluster were necessary.

The attribute pairs displayed are the ones that best correlate according to Pearson's correlation. The label obtained provides, as an aid to the specialist, the interpretation that:

- Cluster 0 is composed of plants whose sepal length (SL) varies between 4.3 cm and 6.1 cm, and the sepal width (SW) varies between 2 cm and 3.2 cm;
- Cluster 1 is formed by plants whose petal length (PL) varies between 0.99 cm and 3.95 cm, and petal width (PW) varies between 0.1 cm and 1.3 cm;
- Cluster 2 is composed of plants whose sepal length (SL) varies between 6.1 cm and 7.9 cm and petal width (SW) varies between 1.3 cm and 2.5 cm;
- Cluster 3 is composed of plants whose petal length (PL) varies between 3.95 cm and 6.9 cm and sepal width (SW) varies between 3.2 cm and 4.4 cm;

It was verified, therefore, that after executing the proposed method, fewer values are included as the range of values decreases, resulting in a non-overlap between labels generated by the same attribute in different clusters. This fact also occurred with the other data sets used.

The Table XIV displays the result of labeling the method of [20] for the Iris dataset. It was observed that the result was generated after 568 iterations of the method based on degrees of membership, which corresponds to a very high computational cost, in addition to having 12 database elements that could not be labeled. This model, proposed in this research, generated a maximum of six iterations to form labels, corroborating and reducing the computational effort. No criteria were used to infer the optimal number of groups in [20], so the author used K=3 for clustering.

TABLE XIV. Rótulos Gerados Por [20], Considerando a Base De Dados Iris

Cluster	Elements	Label		Analysis
		Attributes	Range	Hits(%)
1	50	PW	0.1 ∼0.6	100
		PL	1 ∼1.9	100
2	52	PL	3.5 ∼5.0	82.69
3	36	PL	5.1 ∼6.9	91.66

Table XV compares the models for the Iris dataset, considering the average hit rate of the labels, the number of attribute-value range pairs, and the maximum number of iterations that compose them.

TABLE XV. Comparison between the Labeling Models, Considering the Iris Database

Model	(%)	Attribute-Range Pairs	Maximum Iterations
Model of [20]	91.45	4	568
Proposed Model	97.5	8	6

In this proposed model, it was observed that the computational cost spent on forming the labels was extremely low, favoring a minimum number of iterations, compared to the [20] model, which presented a lower rate than this research model. Using the proposed method for inferring the optimal amount of *clusters* favored the obtained result. Also, in both methods, no range of attribute values overlapped for the same cluster when considering the same attribute.

B. Wine Dataset

Table XVI displays the analysis result for the Wine data set labeling for K=4 and FX=3 according to the proposed method. The clusters were labeled by alcohol and proline for clusters 0, 1, and 2 and by total phenols and flavanoids for cluster 3, as they are the attributes that best correlate according to the correlation coefficient, with a value of 0.86 for alcohol and proline and 0.64 for WK and LK.

TABLE XVI. Wine Database Labeling Analysis

Cluster	Elements	Label		Analysis	Iterations
		Attributes	Range	(%)	
0	66	Alcohol	12.72 ∼13.5	86.76	8
		Proline	690 ∼937	100	0
1	23	Alcohol	13.51 ∼14.83	86.60	13
		Proline	970 ∼1680	100	0
2	57	Alcohol	11.03 ∼12.7	91.22	7
		Proline	278 ∼590	100	0
3	32	Total phenols	0.98∼1.42	100	0
		Flavanoids	0.34∼1.25	100	0

The clusters were labeled by the attributes Alcohol and Proline and also Total phenols and Flavanoids, as they are the ones that best correlate according to the correlation coefficient and also have the highest rates. The interpretation given by the labels is that:

- Cluster 0 is composed of alcohol between 12.72 and 13.5 and proline between 690 and 937;
- Cluster 1 is composed of alcohol between 13.51 and 14.83 and proline between 970 and 1680;
- Cluster 2 is composed of alcohol between 11.03 and 12.7 and proline between 278 and 590;
- Cluster 3 comprises total phenols between 0.98 and 1.42 and flavanoids between 0.34 and 1.25.

The Table XVII shows the result of labeling the method of [14] for the Wine dataset. No criteria were used to infer the optimal number of groups in lopes, so the author used K=3 for clustering and FX=3 for data labeling.

TABLE XVII. Labels Generated by [14], Considering the Wine Database

Cluster	Elements	Label		Analysis
		Attributes	Range	(%)
0	62	Proline	628.5∼979	85.48
1	47	Proline	979∼1680	97.87
2	69	Proline	278∼628.5	100

Table XVIII compares the models in the Wine dataset, considering the average hit rate and the number of attribute-value range pairs that compose them.

It was found that the proposed model has a higher average hit rate than the [14] model, in addition to not having any overlap between ranges of values. This result was favored using the proposed method to infer the optimal number of *clusters*.

TABLE XVIII. COMPARISON BETWEEN THE LABELING MODELS, CONSIDERING THE WINE DATASET

Template	Wine	
	Average Hit Rate (%)	Attribute-Value Range Pairs
Model of [14]	94.45	3
Proposed Model	95.57	6

C. Seeds Dataset

Table XIX shows the analysis result for labeling the Seeds dataset for K=3 and FX=2 according to the proposed method. The clusters were labeled by perimeter (P) and area (A) for clusters 0 and 1 and by seed width (WK) and seed length (LK) for cluster 2, as these are the attributes that best correlate accordingly, with the correlation coefficient, with a value of 0.99 for P and A and 0.86 for WK and LK.

TABLE XIX. SEED DATABASE LABELING ANALYSIS

Cluster	Elements	Label		Analysis	Iterations
		Attributes	Range	(%)	
0	72	P	12.41 ∼13.78	91.04	7
		A	10.59 ∼13.07	89.56	9
1	61	P	13.82 ∼15.33	87.80	8
		A	13.19 ∼16.44	87.80	9
2	77	WK	3.465∼4.033	100	0
		LK	5.826∼6.675	100	0

The attribute pairs displayed are the ones that best correlate according to Pearson's correlation. The following interpretations can be drawn from the dataset labels:

- In cluster 0, elements have a perimeter (P) from 12.41 cm to 13.78 cm and an area (H) from 10.59 cm^2 to 13.07 cm^2;
- In cluster 1, the elements have a perimeter (W) from 13.82 cm to 15.33 cm and an area (H) from 13.19 cm^2 to 16.44 cm^2;
- In cluster 2, elements have seed width (WK) from 3465 to 4033 and seed length (LK) from 5826 to 6675.

The Table XX presents the result of labeling the method of [14] for the Seeds dataset. No criteria were used to infer the optimal number of groups in [14], so the author used K=3 for clustering and FX=3 for data labeling.

TABLE XX. LABELS GENERATED BY [14], CONSIDERING THE SEEDS DATABASE

Cluster	Elements	Label		Analysis
		Attributes	Range	Hits (%)
0	67	A	12.78 ∼16.14	87.30
		P	13.73 ∼15.18	
1	82	P	12.41 ∼13.73	86.58
		A	10.59 ∼12.18	
2	61	P	15.18 ∼17.25	98.34
		A	16.14 ∼21.18	
		LK	5.826 ∼6.675	
		WK	3.465 ∼4.033	

Table XXI addresses a comparison between models for the Seeds dataset, considering the average hit rate and the number of attribute-value range pairs that compose them.

The average hit rate of this proposed model is higher than the [14] method. No overlapping of the range of attribute values was verified in this model under analysis, in addition to having generated more accurate labels. It was observed that this result was favored due to the proposed method for inferring the optimal number of *clusters*.

TABLE XXI. COMPARISON BETWEEN THE LABELING MODELS, CONSIDERING THE SEEDS DATABASE

Model	Iris	
	Average Hit Rate(%)	Attribute-Range Pairs
Model of [14]	90.74	8
Proposed Model	92.70	6

D. Breast Cancer Dataset

Table XXII presents the analysis result for labeling the Breast Cancer data set for K=3 and FX=2 according to the proposed method. The clusters were labeled by Uniformity of Cell Size (UCS) and Uniformity of Cell Shape (UCSH) for cluster 0, Brand Chromatin (BC) and Uniformity of Cell Size (UCS) for cluster 1 and Brand Chromatin (BC) and Uniformity of Cell Shape (UCSH) for cluster 2, as they are the attributes that best correlate according to the correlation coefficient, with a value of 0.91 for UCSH and UCS, 0.76 for BC and UCS and 0.74 for BC and UCSH.

TABLE XXII. BREAST CANCER DATABASE DATA LABELING ANALYSIS

Cluster	Elements	Label		Analysis	Iterations
		Attributes	Range	Hits (%)	
0	455	UCSH	1 ∼4	97.40	4
		UCS	1 ∼2	82.18	3
1	108	BC	1 ∼4	100	0
		UCS	3 ∼10	98.68	2
2	120	BC	5∼10	100	0
		UCSH	5∼10	96.51	5

Although it is a large database with many attributes, the dispersion of the group elements about the dataset's average is small, which generated a few iterations about the Iris, Wine, and Seeds datasets. The label obtained provides the interpretation that:

- Cluster 0 is composed of elements whose Cell Shape Uniformity (UCSH) varies between 1 and 4 and Cell Size Uniformity (UCS) varies between 1 and 2;
- Cluster 1 is composed of elements whose Soft Chromatin (BC) varies between 1 and 4, and Cell Size Uniformity (UCS) varies between 3 and 10;
- Cluster 2 is made up of elements whose Soft Chromatin (BC) ranges from 5 to 10 and Cell Shape Uniformity (UCSH) ranges from 5 to 10.

The Table XXIII shows the result of labeling the method of [21] for the Breast Cancer dataset. No criteria were used to infer the optimal number of groups in [21], so the author used K=2 for clustering.

TABLE XXIII. LABELS GENERATED BY [21], CONSIDERING THE BREAST CANCER DATABASE

Cluster	Elements	Label		Analysis
		Attributes	Range	Hits(%)
0	232	UCS	1 ∼ 5	99.14
		MA	1 ∼ 10	99.14
		BN	1 ∼ 5.97	99.14
1	451	SECS	2 ∼ 10	99.11
		UCS	1.9 ∼ 10	99.11

Table XXIV presents a comparison between the models of the Breast Cancer data set, considering the average hit rate of the labels and the number of attribute-value range pairs that compose them.

TABLE XXIV. Comparison between the Labeling Models, Considering the Breast Cancer Database

Model	Iris	
	Average Hit Rate(%)	**Attribute-Range Pairs**
Model of [21]	99.13	5
Proposed Model	95.79	6

Despite a slightly lower average hit rate, the proposed model does not have overlapping labels, considering the same attribute. It was observed that in the model proposed by [21], there is an overlapping range of values, which compromises the interpretation of the label since the same label referring to the UCS attribute belongs to more than one *cluster*, that is is, UCS(c_1=[1∼5]) and UCS(c_2=[1.9∼10]).

Considering the four sets of data presented, it was verified that the proposed labeling approach does not offer any overlap between ranges of values of the same data set, considering the same attribute. In addition, the number of iterations was greatly reduced, favoring a low computational cost. It was also found in this proposed model that some ranges of values of certain attributes did not require iterations, given the lack of overlap between labels, when considering the same attribute. In three (Iris, Wine, Seeds) of the four data sets compared, the hit rate for the proposed labeling was higher, considering the use of the proposed method for inferring the optimal amount of *clusters* favored this result. The method considered in this paper is free of errors or biases.

VI. Conclusion

The group inference method developed in this research work proved to be satisfactory, considering that it was able to display an optimal number of clusters correlating the value of K to the range of attribute values, contributing to improving the data grouping process about other criteria existing in the literature separately, such as Elbow, Silhouette Coefficient, and Calinski-Harabasz Criterion.

Through labeling, this work provided an improved approach for group interpretation capable of automatically labeling data without overlapping any range of values in the same dataset, considering the same attribute and still with a reduced computational effort. This study initially used four data sets obtained from the UCI Repository, including Iris, Wine, Seeds, and Breast Cancer.

The results obtained in the experiments showed that the approach contributes to the groups' interpretation. The standard deviation-based labeling model also generated satisfactory results, with an average hit rate above 92% for the data sets. The model guarantees an improvement in the specificity of the labels, reducing the computational effort to generate them compared to other methods proposed in the literature.

References

[1] De Araujo, F. N., Machado, V. P., Soares, A. H.,& de MS Veras, R. (2018, July). Automatic cluster labeling based on phylogram analysis. In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

[2] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37-37.

[3] Russell, S. and Norvig, P. (2021). Artificial intelligence: A modern approach, fourth edition.

[4] Manning, C. D., Raghavan, P., & Schütze, H. (2009). Probabilistic information retrieval. Introduction to Information Retrieval, 220-235.

[5] Anaya-Sánchez, H., Pons-Porrata, A., & Berlanga-Llavori, R. (2008). A new document clustering algorithm for topic discovering and labeling. In Progress in Pattern Recognition, Image Analysis and Applications: 13th Iberoamerican Congress on Pattern Recognition, CIARP 2008, Havana, Cuba, September 9-12, 2008. Proceedings 13 (pp. 161-168). Springer Berlin Heidelberg.

[6] Di, J., & Gou, X. (2018). Bisecting K-means Algorithm Based on K-valued Selfdetermining and Clustering Center Optimization. J. Comput., 13(6), 588-595.

[7] Kingrani, S. K., Levene, M., & Zhang, D. (2018). Estimating the number of clusters using diversity. Artificial Intelligence Research, 7(1), 15-22.

[8] Zhou, S., Xu, Z., & Liu, F. (2016). Method for determining the optimal number of clusters based on agglomerative hierarchical clustering. IEEE transactions on neural networks and learning systems, 28(12), 3007-3017.

[9] MacQuuen, J. B. (1967). Some methods for classification and analysis of multivariate observation. In Proceedings of the 5th Berkley Symposium on Mathematical Statistics and Probability (pp. 281-297).

[10] Pinheiro, J., Cunha, S., Gomes, G., and Carvajal, S. (2013). Probabilidade e estatística: quantificando a incerteza. Elsevier Brasil.

[11] Bland, J. M. and Altman, D. G. (1996). Statistics notes: measurement error. Bmj, 312(7047):1654.

[12] MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA.

[13] Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics-theory and Methods, 3(1), 1-27.

[14] Lopes, L. A., Machado, V. P., Rabêlo, R. A., Fernandes, R. A., & Lima, B. V. (2016). Automatic labelling of clusters of discrete and continuous data with supervised machine learning. Knowledge-Based Systems, 106, 231-241.

[15] Gustriansyah, R., Suhandi, N., & Antony, F. (2020). Clustering optimization in RFM analysis based on k-means. Indonesian Journal of Electrical Engineering and Computer Science, 18(1), 470-477.

[16] Solikhun, S., Yasin, V., & Nasution, D. (2022). Optimization of the Number of Clusters of the K-Means Method in Grouping Egg Production Data in Indonesia. International Journal of Artificial Intelligence & Robotics (IJAIR), 4(1), 39-47.

[17] Catlett, J. (1991). On changing continuous attributes into ordered discrete attributes. In Machine Learning—EWSL-91: European Working Session on Learning Porto, Portugal, March 6–8, 1991 Proceedings 5 (pp. 164-178). Springer Berlin Heidelberg.

[18] Hwang, G. J., & Li, F. (2002). A dynamic method for discretization of continuous attributes. In Intelligent Data Engineering and Automated Learning—IDEAL 2002: Third International Conference Manchester, UK, August 12–14, 2002 Proceedings 3 (pp. 506-511). Springer Berlin Heidelberg.

[19] Thirumalai, C., Chandhini, S. A., and Vaishnavi, M. (2017). Analysing the concrete compressive strength using pearson and spearman. In 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), volume 2, pages 215–218. IEEE.

[20] Imperes Filho, F., Machado, V. P., Veras, R. d. M. S., Aires, K. R. T., and Silva, A. M. L. (2020). Group labeling methodology using distance-based data grouping algorithms. Revista de Informática Teórica e Aplicada, 27(1):48–61.

[21] Silva, L. E. S., Machado, V. P., Araujo, S. S., Lima, B. V. A. d., and Veras, R. d. M. S. (2021). Using regression error analysis and feature selection to automatic cluster labeling. In EPIA Conference on Artificial Intelligence, pages 376–388. Springer.